Interspeech 2022
18-22 September 2022, Incheon, Korea

Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation

Dan Berrebbi[1], *Jiatong Shi*[1], *Brian Yan*[1], *Osbel López-Francisco*[2], *Jonathan Amith*[3], *Shinji Watanabe*[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]Universidad Nacional Autónoma de México, Iztacala
[3]Dept. of Anthropology, Gettysburg College

dberrebb@andrew.cmu.edu, jiatongs@andrew.cmu.edu

Abstract

Self-Supervised Learning (SSL) models have been successfully applied in various deep learning-based speech tasks, particularly those with a limited amount of data. However, the quality of SSL representations depends highly on the relatedness between the SSL training domain(s) and the target data domain. On the contrary, spectral feature (SF) extractors such as log Mel-filterbanks are hand-crafted non-learnable components, and could be more robust to domain shifts. The present work examines the assumption that combining non-learnable SF extractors to SSL models is an effective approach to low resource speech tasks. We propose a learnable and interpretable framework to combine SF and SSL representations. The proposed framework outperforms significantly both baseline and SSL models on Automatic Speech Recognition (ASR) and Speech Translation (ST) tasks on three low resource datasets. We additionally design a mixture of experts based combination model. This last model reveals that the relative contribution of SSL models over conventional SF extractors is very small in case of domain mismatch between SSL training set and the target language data.

Index Terms: Low Resource, Self-Supervised Learning, Spectral Features, co-Attention, Mixture of Experts.

1. Introduction

End-to-end models based on deep learning have demonstrated their superiority over conventional hidden Markov-based models on speech tasks for some corpora [1–4]. End-to-end models could be beneficial to low resource speech tasks because these models: (1) alleviate the need of language specific resources such as lexicons [5–7]. (2) can be trained multilingually to facilitate cross-lingual transfers between high resource and low resource languages through shared architecture and weights [8]. On the other hand, end-to-end models can perform poorly when the training data is limited [9] and low resource scenarios often introduce a language-mismatch with the data used to train powerful self-supervised learning (SSL) representations [10].

One direction towards mitigating these low-resource issues is to incorporate knowledge from several languages into multilingual end-to-end models [11–13]. When there is no training data available for the target languages, these systems can be even applied in a zero-shot manner [14–16]. Fortunately, many languages have small amounts of data which can be used to fine-tune large-scale multilingual models towards target languages, resulting in further improvements [17–20].

Another direction is to use self-supervised learning models trained on large untranscribed corpora as front-end feature extractors, replacing conventional spectral features (SF) such as log Mel-filterbanks coefficients (FBANK) [21–26]. During their unsupervised training, SSL models [27–30] learn their own feature extraction modules and are totally free of SF at fine-tuning time. As these models achieve state of the art on numerous speech tasks and significantly outperform models with more supervision, the effectiveness of SF on low resource tasks is increasingly questioned.

The majority of SSL models are trained exclusively using English speech. Although these approaches have shown improvements, even when domain mismatches occur (such as language or audio conditions [31]), performance depends on the relatedness between the SSL training domain and the target language one [32]. SSL first layers output representations tend to be quite similar to SF according to a canonical correlation analysis [33] of Wav2vec2 [29] from Pasad et al. [34]. In contrast, the last layers are likely to be more corpus or domain-specific, which should be randomly initialized at fine-tuning time [34]. Therefore, we assume that SSL representations are potentially more hurted by domain shifts than SF-based systems are. SF are domain and language agnostic and their use in multilingual models has demonstrated that they enable strong cross-lingual transfers [8]. It is then legitimate to assume that a model leveraging both SF and SSL representations would lead to strong performances on low resource speech scenarios.

In the present work, we examine this assumption by building a framework that enables combining SF and SSL representations through learnable fusions. We propose linear, convolutional and co-attention based combinations. Those methods obtain a relative diminution of 19.3% Character Error Rate (CER), averaged on two ASR datasets, and a gain of 1.0 BLEU, on an ST dataset, over the SSL baseline model, while having less than 0.01% additional parameters. We further propose a mixture of experts [35] based technique in order to better interpret the roles and complementarities of SF and SSL components.[1] Finally the proposed framework is evaluated on Totonac, a Mexican endangered language, and we release the first publicly available annotated speech corpus of this language.[2]

2. Speech Representations

Spectral Features: Machine learning based speech analytics require the extraction of feature vectors from raw analog waveforms. Log Mel-filterbanks features (FBANK), conventionally used for supervised speech processing tasks, are perceptually inspired by human hearing. These features sample and quan-

[1]Our code is released on ESPnet [36]
[2]http://www.openslr.org/107/

10.21437/Interspeech.2022-10796

tize the analog waveform, apply pre-emphasis to boost high frequency energies, undergo a discrete Fourier transform (DFT), and finally passed through Mel filter banks. It is worth noting that the DFT operation is linear and could be learned during model training but the system may fail to learn it due to its high complexity, especially if only small amounts of data are available.

Self-Supervised Learning features : While FBANK are handcrafted features inspired by the human perception of speech, SSL features learn latent representations derived from large amounts of unlabeled data. After training the SSL model, often referred to as pre-training, a fine-tuning phase is conducted with a task-specific labeled data set. The key idea is that unlabeled data contains valuable information and is far more abundant than labeled data in any domain. This paradigm leads to general-purpose speech representation, suitable for speech processing tasks [10].

3. Proposed Approaches

3.1. Feature extraction

Let S be a sampled and quantized raw waveform of one utterance. We note $f_{\text{SF}}(S)$ and $f_{\text{SSL}}(S)$ the features extracted from S by spectral feature extractors and SSL models (respectively SF and SSL in formulas). We note T_{SF} and T_{SSL} the number of frames of the utterance, while D_{SF} and D_{SSL} denote dimensions of the features extracted by f_{SF} and f_{SSL}. We obtain,

$$f_{\text{i}}(S) = (f_{\text{i}}^{t}(S) \in \mathbb{R}^{D_{\text{i}}} | t = 1, \cdots, T_{\text{i}}),\ i \in \{\text{SF, SSL}\} \quad (1)$$

Additional linear projection and reshaping is applied over SF and SSL features to allow a same feature dimension $D = D_{\text{SF}}$ and number of frames $T = T_{\text{SSL}}$. For the dimension, we choose to project SSL features into SF space and not the inverse in order to decrease the number of parameters (as $D_{\text{SF}} < D_{\text{SSL}}$) for efficiency purposes. For the number of frames, as we use a frame-shift two times longer for SSL than for SF, we downsample (through linear projection and reshaping) the SF features to get a common number of frames $T = T_{\text{SSL}}$. We now have $f_{\text{SF}}(S) \in \mathbb{R}^{T \times D}$ and $f_{\text{SSL}}(S) \in \mathbb{R}^{T \times D}$. Our goal is to combine $f_{\text{SF}}(S)$ and $f_{\text{SSL}}(S)$ in order to get the best model for low resource tasks.

3.2. Learnable combinations

We first propose a general framework of using learnable transformations (concatenation, convolutional, and co-attention [37] mechanisms) for combining those features. Such learnable fusions have previously been employed in various multi-source/multimodal applications [38, 39]. The framework is formulated as follows, where $f_{\text{FUSE}}(S)$ is the resultant features:

$$f_{\text{FUSE}}(S) = \textsc{Linear}(\textsc{Transform}(f_{\text{SF}}(S), f_{\text{SSL}}(S)) \quad (2)$$

With TRANSFORM being a concatenation, a convolution or a **co-attention based fusion**. We will dive into more details about Eq. (2) for the proposed co-attention fusion method, which is illustrated in Fig. 1.

Let W_{SF}^{Q}, W_{SF}^{K}, W_{SF}^{V}, $W_{\text{SSL}}^{\text{Q}}$, $W_{\text{SSL}}^{\text{K}}$ and $W_{\text{SSL}}^{\text{V}}$ be six learnable matrices of $\mathbb{R}^{D \times D}$. We use classical attention notation [40] in Eq. (3). For i$\in$\{SF, SSL\}, we note,

$$Q_{\text{i}} = f_{\text{i}}(S) W_{\text{i}}^{\text{Q}}, \quad K_{\text{i}} = f_{\text{i}}(S) W_{\text{i}}^{\text{K}}, \quad V_{\text{i}} = f_{\text{i}}(S) W_{\text{i}}^{\text{V}} \quad (3)$$

Then, we apply two cross-attention blocks in parallel, each made of a one head scaled dot-product attention operation, with

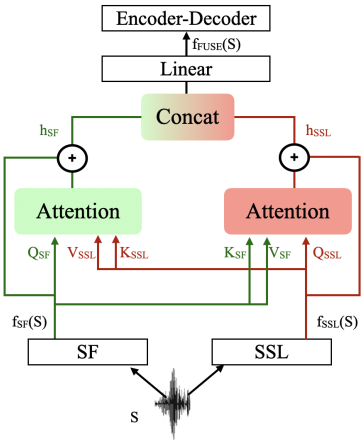

Figure 1: *Architecture of our proposed co-attention based fusion. Raw signal S is passed through SF and SSL feature extractors. The extracted features, $f_{SF}(S)$ and $f_{SSL}(S)$, attend to each other through two distinct attention mechanisms. Output features are then concatenated, projected and passed to the speech model.*

residual connection. We obtain the SF context vector h_{SF} by using the SF feature vector as a query and the SSL feature vector as key and value, and vice versa to obtain h_{SSL}, the SSL context vector. Eq. (4) and Eq. (5) describe those symetric attention mechanisms, where $\cdot$ is the dot-product operator.

$$h_{\text{SF}} = \textsc{SoftMax}(\frac{Q_{\text{SF}} \cdot K_{\text{SSL}}}{\sqrt{D}}) V_{\text{SSL}} + f_{\text{SF}}(S) \quad (4)$$

$$h_{\text{SSL}} = \textsc{SoftMax}(\frac{Q_{\text{SSL}} \cdot K_{\text{SF}}}{\sqrt{D}}) V_{\text{SF}} + f_{\text{SSL}}(S) \quad (5)$$

Our final feature is a projection on $\mathbb{R}^{D}$ of the concatenation of h_{SF} and h_{SSL}, as descibed in Eq. (6), where $\|$ design the vector concatenation operation.

$$f_{\text{FUSE}}(S) = \textsc{Linear}(h_{\text{SF}} \,\|\, h_{\text{SSL}}) \quad (6)$$

We also designed an attention-based fusion, however performance on preliminary experiments were weak compared to the co-attention model. We assume that the parallel computations on SF and SSL enable more sophisticated combinations of the two feature extractors than only one attention block would do.

3.3. Mixture of Experts

To get a broader understanding of the potential complementarity of SF and SSL features, we propose an adaptation of the mixture of experts [35] gating paradigm, illustrated in Fig. 2. We consider the two feature extractors, f_{SF} and f_{SSL}, as our experts. This model requires a same number of frames for the two experts (see the processing step in Sec. 3.1). We use $f_{\text{SF}}(S)$ as input feature to the gate.[3] Weights are calculated following Eq. (7), where $w(S) \in \mathbb{R}^{T \times 2}$ is the obtained weight matrix.

$$w(S) = \Theta(f_{\text{SF}}(S) W_{\text{MoE}}), \quad (7)$$

with $W_{\text{MoE}} \in \mathbb{R}^{D \times 2}$ a learnable matrix, and $\Theta(\cdot)$ a gating-type function such as SOFTMAX. For clarity, we introduce $w_{\text{SF}}(S)$, $w_{\text{SSL}}(S) \in \mathbb{R}^{T}$, the column vectors of $w(S)$.

The final combined feature is computed following Eq. (8), where $[x]^{\text{tr}}$ denotes the transpose vector of x.

$$f_{\text{FUSE}}(S) = \sum_{i \in \{\text{SF,SSL}\}} [w_{\text{i}}(S)]^{\text{tr}} f_{\text{i}}(S) \quad (8)$$

[3] Both $f_{\text{SF}}(S)$ or $f_{\text{SSL}}(S)$ could be used as input for the gate layer. We discuss this designing choice in Sec 4.2.

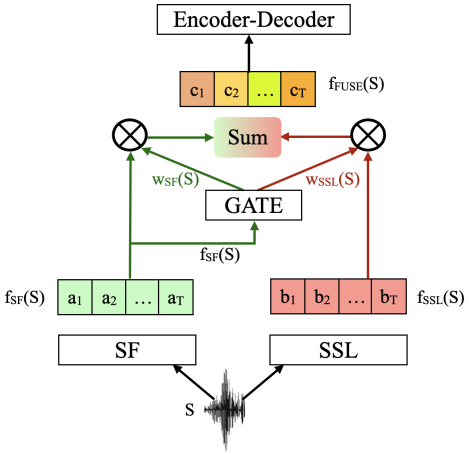

Figure 2: *Architecture of the model combining SF and SSL through a gating mechanism. For a given utterance, the features are extracted by the two models (a_i for SF and b_i for SSL, $i \in \{1, ..., T\}$). Each model gets confidence scores and features are then summed. The c_i variables indicates the weighted sum. Colors of c_i frames are used to show how each frame gets a specific combination of SF (green) and SSL (red) features.*

The mixture of experts model outputs a weighted sum of feature extractors for each frame of the utterance. The weights can be interpreted as confidence scores of SF and SSL for each frame. This model makes the fusion process more interpretable by enabling to compare relative usage of SF and SSL.

4. Experiments

4.1. Datasets

Totonac is an endangered language spoken in the northern sierras of the state of Puebla and adjacent areas of Veracruz, Mexico. To increase the coverage over endangered languages, we evaluate our proposed methods on Totonac and release a publicly available version of Totonac ASR data.[4] The corpus comprises 10 hours of speech (86 long recordings) with fine-grained transcriptions. We randomly selected 70 recordings for the training set, 8 for validation, and 8 for testing.[5] In addition to Totonac, we perform experiments on Arabic corpora of 20 hours from Commonvoice 5.1 [41], still in the low-resource scenario. Finally, we extend our study to low resource ST using the Mboshi-French dataset [42], consisting of 4 hours of speech, to show that our framework is effective in other speech tasks as well. We chose Arabic (Semitic language) and Mboshi (Bantu language) as they belong to different language groups than English (Germanic). Thus, we will compare the robustness of the SSL representations to the ones of our proposed models over a set of diverse language families, all different from the one of the SSL self-training data.

4.2. Experimental setup

Baseline : Our ASR baseline (**Base** in the experiments) adopts a transformer-based encoder-decoder architecture with CTC/Attention hybrid training [43]. The front-end extracts FBANK spectral features with a frame length of 25ms and a frame-shift of 10ms. The extracted FBANK features are subsampled with a convolutional block and then fed into the encoder-decoder. The encoder consists of 12 self-attention blocks with 4-head attention and 256-dimensional hidden sizes while the decoder has 6 cross-attention transformer blocks. For ST, we add 2 extra decoders of 2 layers each to this architecture. SpecAugment [44] and speed perturbation are employed for data augmentation. Hyperparameters used for training can be found on ESPnet. The ASR model is trained to recognize 250 byte-pair-encoding (BPE) units. The same architecture and training configuration are used for the following experiments.

Self-supervised representations : In our experiments, we employ HuBERT [27] , which shows promising results over the SUPERB benchmark [10].[6] To fully explore the potential of HuBERT, we select the HuBERT-large model pre-trained over 60k hours of LibriLight [46, 47]. The SSL wrapper provided in Yang et al. [48] is applied to extract high-dimensional features with a 20ms frame-shift. In experiment **SSL**, the FBANK feature extractor (used in **Base**) is replaced by the pretained HuBERT model, which is fine-tuned during training.[7]

Learnable combinations : Experiments **Linear, Conv.** and **co-Att.** are the TRANSFORM operations introduced in Eq. (2) of Sec 3.2 respectively for concatenation, convolutional, and co-attention based fusions. For **Linear** experiment, we concatenate $f_{\text{SF}}(S)$ and $f_{\text{SSL}}(S)$ and then project the concatenation into a 80-dimensional space. In **Conv.** experiment, we apply a 1-dimensionnal convolutional layer with kernel size 5 and stride of 1 over $f_{\text{SF}}(S)$ and $f_{\text{SSL}}(S)$ before concatenating and projecting them. The co-attention model is described through Eq. (3) to Eq. (6), and the model is illustrated in Fig. 1.

Mixture of experts : Our mixture of experts model (**MoE** in the experiments) follows Eq. (7) and Eq. (8) described in Sec. 3.3. For the main experiments, we use SF (here FBANK) as input features and $\Theta(\cdot) = \text{LOG-SOFTMAX}(\cdot)$ for the gating function. We performed a comparative study of inputs to the gating function. Using both SF or SSL features led to better scores than the baselines but SF as input performed best. Our interpretation is that it is easier for the model to learn gating weights when computed over non-learnable features (SF, here FBANK) than over complex features which are continuously fine-tuned. We also compared results with $\Theta(\cdot) = \text{LOG-SOFTMAX}(\cdot)$ and $\Theta(\cdot) = \text{SOFTMAX}(\cdot)$. Performances are similar, $\Theta(\cdot) = \text{LOG-SOFTMAX}(\cdot)$ being slightly better. A more detailed analysis of the gating weights intra-utterance revealed a more peaky behavior for $\Theta(\cdot) = \text{SOFTMAX}(\cdot)$, which in our opinion led to the small performance degradation.

Evaluation metrics : We use Character Error Rate (CER) for evaluation of our ASR models and BLEU score to measure performances of our ST systems.

5. Results and Analysis

5.1. Main results

Table 1 provides results for the experiments listed in Sec. 4.2.

5.1.1. Speech Recognition results

First we remark that using HuBERT as a feature extractor (**SSL** experiment) instead of FBANK (**Base** experiment) is very ef-

[4] http://www.openslr.org/107/

[5] Those splits are officially released at https://github.com/ftshijt/Totonac_Split.git

[6] We also performed preliminary experiments over Wav2vec2 XLSR model [45], but it did not improve the results over HuBERT model so we continued the study only for HuBERT model.

[7] SSL based front-ends could be freezed, but the best performances were obtained when fine-tuning the models.

Table 1: *ASR and ST results over models described in Sec. 4.2. The two first experiments are our FBANK and SSL baselines. The following lines are the proposed Linear, Convolutional, co-Attention, and Mixture of Experts models.*

	CER ↓		BLEU ↑
Exp	Totonac	Arabic	Mboshi-French
Base	17.2	15.4	10.9
SSL	14.2	8.1	10.6
Linear	14.0	6.6	**11.6**
Conv.	13.9	7.2	11.3
co-Att.	**13.4**	**5.4**	10.9
MoE	13.7	6.2	11.2

fective on the Totonac and Arabic ASR corpora, leading to respective diminutions of 3.0 and 7.3 of CER.

Then, we note that all of the combination methods we introduced in Sec. 3.2 led to improvements on the two datasets over the **Base** and **SSL** models. We get a diminution of 2.7 CER (33%) on Arabic and 0.8 CER (5.6%) on Totonac when using the co-attention model. The co-attention model performs better than the linear and convolution based methods, in particular for Arabic. A possible explanation is that this model: (1) has a larger modeling capacity (leading to better results), and (2) induces a more balanced use of the two front-ends, through the symmetric architecture and the residual connections. This second point could explain the greater CER reduction over Arabic than Totonac, as an equal contribution of front-ends seems to lead to a robust model for Arabic (see Sec. 5.2).

Finally, the mixture of experts model that we introduced for gaining interpretability is also getting strong performances.

5.1.2. Speech Translation results

As it is straightforward to use our front-end fusion framework for different speech tasks, we applied it to ST. Table 1 shows that all of our proposed methods outperforms both FBANK and SSL baselines. We note that using HuBERT representations as front-end degraded the performance in that scenario (see experiments **Base** and **SSL**). Even in that case, all the proposed systems performed better than both baselines. The linear fusion method reaches a BLEU score of 11.6, gaining 1.0 BLEU over the SSL baseline and 0.7 BLEU over the FBANK one. Contrary to the ASR scenario, here the linear and convolutional methods outperform the co-attention one. As Mboshi is only made of only 4 hours of speech, we assume that the co-attention model may be too complex to be well trained contrary to the linear model.

5.2. Mixture of Experts : Weights and Analysis

In this section, we examine the weights $w_{\text{SF}}(S)$ and $w_{\text{SSL}}(S)$ (introduced in Sec. 3.3) obtained by the mixture of experts model for the two ASR datasets. For more interpretability, we normalized them so that $w_{\text{SSL}}(S) + w_{\text{SF}}(S) = 1$. First we can note that our robust **MoE** model is indeed using both FBANK and HuBERT components as the two weights are non negligible. Then, we remark that the weights across frames of a same utterance are quite similar. The two front-ends are used consistently over the frames, which we would expect as a utterance content may be quite consistent. We note that the weights across different utterances are also similar within languages. However they are very different from one language to another.

The first column of Table 2 presents the mean $w_{\text{SSL}}(S)$ weight for each language. Contrary to the Arabic model, which uses HuBERT and FBANK with similar weights, the Totonac model seems to be using HuBERT representations as an adjustment component, relying on average at more than 80% on spectral features. Our interpretation is that the Commonvoice Arabic data is closer in domain (read speech) to the English Libri-Light than Totonac data is (spontaneous speech/conversation). For that reason, HuBERT model may extract relatively better speech representations (compared to FBANK representations) for Arabic than it does for Totonac. This would explain that the mixture of experts model grants HuBERT with a larger weight for the Arabic data. Another way of quantifying HuBERT representations quality over the languages could be to calculate the character error reduction rate (CERR[8]) between FBANK and HuBERT baselines (experiments **Base** and **SSL** in Table 1). The second column in Table 2 confirms our intuition : the mixture of experts model weights the components according to their relative strength over the language. As Arabic benefits more from HuBERT representations than Totonac does, the mixture of experts model assigned a higher weight to the HuBERT front-end in the Arabic model than in the Totonac one.

Table 2: *Two views on HuBERT representations quality over Totonac and Arabic data. The first column presents $\overline{w_{SSL}(S)}$, the mean **MoE** weights for HuBERT front-end. The second column is the character error reduction rate reduction (CERR[8]) between the FBANK baseline and the HuBERT baseline.*

Language	$\overline{w_{\text{SSL}}(S)}$	CERR(**Base** → **SSL**)
Totonac	0.17	17%
Arabic	0.51	47%

6. Conclusions

SSL models performance depends highly on the relatedness between the self-supervised training domain(s) and the target data domain. As spectral features are not subject to those variations, we proposed a framework to combine spectral features to SSL representations. This framework can be applied to many speech tasks with no further work. We obtained strong improvements over ASR and ST datasets compared with the SSL baseline. We further proposed a weight analysis showing that: (1) our models performances are strong for both in-domain and out-of-domain scenarios. (2) our mixture of experts framework enables quantifying the domain shift between the SSL training data and the target language resources.

Future work could involve fusions at the encoder level. As SSL models also perform strongly when used as encoders, fusing SSL features with SF passed through a pre-trained encoder could be an even more robust technique.

7. Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) [49], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system [50], as part of project cis210027p, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

[8]CERR is defined as follows : CERR = $\frac{\text{CER(Base) - CER(SSL)}}{\text{CER(Base)}} \times 100$.

8. References

[1] C.-C. Chiu, T. N. Sainath, Y. Wu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[2] S. Karita, Chen *et al.*, "A comparative study on transformer vs RNN† in speech applications," in *ASRU*, 2019.

[3] N.-Q. Pham, T.-S. Nguyen *et al.*, "Very deep self-attention networks for end-to-end speech recognition," *Interspeech*, 2019.

[4] P. Guo, F. Boyer *et al.*, "Recent developments on ESPnet toolkit boosted by conformer," in *ICASSP*, 2021.

[5] L. A. Grenoble, P. K. Austin, and J. Sallabank, "The Cambridge handbook of endangered languages," *Cambridge University Press*, 2011.

[6] A. Zahrer, A. Zgank, and B. Schuppler, "Towards building an automatic transcription system for language documentation: Experiences from muyu," in *LREC*, 2020.

[7] J. Shi, J. D. Amith, R. Castillo García *et al.*, "Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec," in *EACL*, 2020.

[8] J. Cho, M. K. Baskar *et al.*, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *SLT*, 2018.

[9] C. Lüscher, E. Beck, K. Irie *et al.*, "RWTH ASR systems for Librispeech: Hybrid vs attention - w/o data augmentation," in *Interspeech*, 2019.

[10] H.-S. Tsai, H.-J. Chang, W.-C. Huang *et al.*, "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," *arXiv preprint arXiv:2203.06849*, 2022.

[11] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU*, 2017.

[12] S. Toshniwal, T. N. Sainath *et al.*, "Multilingual speech recognition with a single end-to-end model," in *ICASSP*, 2018.

[13] A. Kannan, A. Datta, T. N. Sainath *et al.*, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," in *Interspeech*, 2019.

[14] X. Li, S. Dalmia, J. Li *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP*, 2020.

[15] B. Yan, S. Dalmia *et al.*, "Differentiable allophone graphs for language-universal speech recognition," *Interspeech*, 2021.

[16] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.

[17] W. Hou, Y. Dong, B. Zhuang *et al.*, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Interspeech*, 2020.

[18] V. Pratap, A. Sriram, P. Tomasello *et al.*, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Interspeech*, 2020.

[19] O. Adams, M. Wiesner *et al.*, "Massively multilingual adversarial speech recognition," in *ACL*, 2019.

[20] B. Li, R. Pang, T. N. Sainath *et al.*, "Scaling end-to-end models for large-scale multilingual ASR," *ASRU*, 2021.

[21] C. Yi, J. Wang, N. Cheng *et al.*, "Applying wav2vec2.0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[22] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-Supervised Representations Improve End-to-End Speech Translation," in *Interspeech*, 2020.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[24] K. D. N, P. Wang, and B. Bozza, "Using Large Self-Supervised Models for Low-Resource Speech Recognition," in *Interspeech*, 2021.

[25] X. Chang, T. Maekaku, P. Guo *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," *ASRU*, 2021.

[26] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *TASLP*, 2021.

[27] W. Hsu, B. Bolte, Y. H. Tsai *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[28] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech*, 2019.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[30] S. Chen, C. Wang, Z. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[31] R. Sanabria, W.-N. Hsu, A. Baevski, and M. Auli, "Measuring the impact of individual domain factors in self-supervised pre-training," *arXiv preprint arXiv:2203.00648*, 2022.

[32] A. Conneau, K. Khandelwal, N. Goyal *et al.*, "Unsupervised cross-lingual representation learning at scale," *Interspeech*, 2019.

[33] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[34] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.

[35] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, 1991.

[36] S. Watanabe, T. Hori, S. Karita *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018.

[37] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016.

[38] J. Libovický and J. Helcl, "Attention strategies for multi-source sequence-to-sequence learning," in *ACL*, 2017.

[39] C. Hori, T. Hori, T.-Y. Lee *et al.*, "Attention-based multimodal fusion for video description," in *EECV*, 2017.

[40] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *NeurIPS*, 2017.

[41] R. Ardila, M. Branson *et al.*, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[42] P. Godard, G. Adda, M. Adda-Decker *et al.*, "A very low resource language speech corpus for computational language documentation experiments," in *LREC*, 2018.

[43] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017.

[44] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019.

[45] A. Conneau, A. Baevski, R. Collobert *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech*, 2021.

[46] J. Kahn, M. Rivière *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.

[47] M. Ott, S. Edunov, A. Baevski *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL*, 2019.

[48] S. Yang, P.-H. Chi, Y.-S. Chuang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[49] J. Towns, T. Cockerill *et al.*, "XSEDE: Accelerating scientific discovery," *Computing in Science & Engineering*, 2014.

[50] N. A. Nystrom, M. J. Levine *et al.*, "Bridges: a uniquely flexible HPC resource for new communities and data analytics," in *XSEDE*, 2015.